10-704 Homework 1
Due: Thursday 2/5/2015

Instructions: Turn in your homework in class on Thursday 2/5/2015

1. **Information Theory Basics and Inequalities** C&T 2.47, 2.29

   (a) A deck of n cards in order $1, 2, \ldots, n$ is given to you. You remove one card at random and then place it again at one of the n available positions at random. What is the entropy of the resulting deck?

   Solution: There are n choices for the card you select and n choices for where the card is placed in the deck. If you choose the ith card and place it back in its original location, then you arrive at the original sequence. Therefore the original sequence occurs with probability $1/n$.

   There are $n-1$ outcomes that each occur with probability $2/n^2$. These are the outcomes where two adjacent items in the list are swapped (i.e. $(2, 1, 3, 4)$).

   The remaining outcomes occur with probability $1/n^2$ and there are $n^2 - n - 2(n-1) = (n-1)(n-2)$.

   The entropy is therefore:

   $$\frac{1}{n}\log n + \frac{2(n-1)}{n^2}\log\frac{n^2}{2} + \frac{(n-1)(n-2)}{n^2}\log n^2$$

   (b) Let X, Y, Z be joint random variables. Prove the following inequalities and identify conditions for equality.

   i. $H(X, Y|Z) \geq H(X|Z)$

   ii. $I(X, Y; Z) \geq I(X; Z)$

   iii. $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$

   iv. $I(X; Z|Y) \geq I(Z; Y|X) - I(Z; Y) + I(X; Z)$

   Solution:

   i.

   $$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \geq H(X|Z)$$

   since entropy is non-negative. This inequality is tight when $H(Y|X, Z)$ is zero, or conditionally on both X, Z the value of Y is deterministic

   ii.

   $$\begin{aligned} I(X, Y; Z) &= H(X, Y) - H(X, Y|Z) = H(X) + H(Y|X) - H(X|Z) - H(Y|X, Z) \\ &= I(X; Z) + H(Y|X) - H(Y|X, Z) \geq I(X; Z) \end{aligned}$$

   The last inequality follows since conditioning cannot reduce entropy. This inequality is tight when $Y \perp Z|X$.

iii. By the chain rule, the left hand side is $H(Z|X,Y)$ while the right hand side is $H(Z|X)$. The inequality follows since conditioning does not reduce entropy. It is tight when $Z \perp Y|X$.

iv. Notice that:

$$\begin{aligned}I(X;Z|Y) - I(Y;Z|X) &= H(Z|Y) - H(Z|X,Y) - H(Z|X) + H(Z|X,Y) \\ &= H(Z|Y) - H(Z|X)\end{aligned}$$

while:

$$I(X;Z) - I(Y;Z) = H(Z) - H(Z|X) - H(Z) + H(Z|Y) = H(Z|Y) - H(Z|X)$$

So this inequality is always an equality.

(c) Consider a distribution on $\{1,\ldots,m\}$ with $\mathbb{P}(X=i) = p_i$. We will assume $p_1 \geq p_2 \geq \cdots \geq p_m$. Let $\mathbf{p} = [p_1,\ldots,p_m]$. Since $X = 1$ is the most likely assignment, the minimal probability of error predictor of X is $\widehat{X} = 1$ with probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$ to find a bound on P_e in terms of the entropy. This is Fano's inequality in the absence of conditioning.

Solution: We will maximize $H(p) = -\sum p_i \log p_i$ subject to $1 - p_1 = P_e$ and $\sum_i p_i = 1$. The first constraint can be handled by substituting $p_1 = 1 - P_e$ and for the second constraint we will optimize the Lagrangian.

$$L(p,\lambda) = -\sum_{i=2}^{m} p_i \log p_i - (1-P_e)\log(1-P_e) + \lambda\left((1-P_e) + \sum_{i=2}^{m} p_i - 1\right)$$

The derivative with respect to p_i $(i \neq 1)$ is:

$$\frac{\partial L(p,\lambda)}{\partial p_i} = 1 - \log p_i + \lambda = 0 \Rightarrow p_i = \exp(1+\lambda)$$

Setting the derivative with respect to λ equal to zero, implies that $\sum_{i=2}^{m} p_i = P_e$, and this means:

$$\sum_{i=2}^{m} \exp(1+\lambda) = P_e \Rightarrow \lambda = \log(P_e/(m-1)) - 1$$

So that $p_i = P_e/(m-1)$. This means:

$$H(p) \leq -(1-P_e)\log(1-P_e) + P_e \log\left(\frac{m-1}{P_e}\right) = H(P_e) + P_e \log(m-1)$$

which is Fano's inequality without conditioning.

2. **Estimation of Entropy Functionals** In class we mentioned that there are no practical unbiased estimators for entropy functionals. One can however design an unbiased estimator if you are allowed to choose a set of samples of arbitrary but finite size. The problem is that there is no *a priori* bound on the sample size. In this question we will develop and analyze these estimators for the discrete setting. Let $X_1, X_2, \dots$ denote a sequence of samples from a discrete distribution P with symbols $C_1, \dots, C_k$ and probabilities $(p_1, \dots, p_k)$.

   (a) For $1 \le i \le k$, let N_i denote the smallest $j \ge 1$ for which $X_j = C_i$. Show that:

$$\widehat{H}_1 = \sum_{i=1}^{k} \frac{\mathbf{1}[N_i \ge 2]}{N_i - 1} \tag{1}$$

   is an unbiased estimator for the entropy $H(P) = -\sum_{i=1}^{k} p_i \log p_i$.

   Solution: Notice that the marginal distribution N_i is a geometric distribution, so that $\mathbb{P}[N_i = j] = p_i(1-p_i)^{j-1}$.

$$\begin{aligned}\mathbb{E}\widehat{H}_1 &= \sum_{i=1}^{k} \mathbb{E}\frac{\mathbf{1}[N_i \ge 2]}{N_i - 1} = \sum_{i=1}^{k}\sum_{j=2}^{\infty} \frac{p_i(1-p_i)^{j-1}}{j-1} \\ &= \sum_{i=1}^{k} p_i \sum_{j=2}^{\infty} \frac{(1-p_i)^{j-1}}{j-1} = \sum_{i=1}^{k} p_i \sum_{j=1}^{\infty} \frac{(1-p_i)^{j}}{j} \\ &= -\sum_{i=1}^{k} p_i \log p_i\end{aligned}$$

   The last line follows from the expansion: $\log(1-x) = -\sum_{j=1}^{\infty} x^j/j$.

   (b) Design an unbiased estimator based on pairing each of the first n samples with the next sample in the sequence with the same symbol. The identity $\frac{\log(1-x)}{1-x} = -\sum_{i=1}^{\infty} h_i x^i$ where $h_i = \sum_{j=1}^{i} \frac{1}{j}$ is the ith harmonic number will be useful.

   Solution: For each of the first n samples $i \in [n]$, let ω_i be the smallest $j \ge i$ such that X_i and X_{j+1} are the same symbol. Define:

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} h_{\omega_i - 1}$$

   By linearity of expectation, it is sufficient to analyze a single term in this summation, say the first term.

$$\begin{aligned}\mathbb{E}\widehat{H} = \mathbb{E}h_{\omega_1 - 1} &= \sum_{i=1}^{k} \mathbb{P}[X_1 = C_i]\mathbb{E}[h_{\omega_1 - 1} | X_1 = C_i] \\ &= \sum_{i=1}^{k} p_i \sum_{j=1}^{\infty} h_{j-1} p_i (1-p_i)^{j-1} = -\sum_{i=1}^{k} p_i \log p_i\end{aligned}$$

This calculation uses the fact that conditional on the symbol of X_1, ω_1 is geometrically distributed. The last step follows from the identity.

(c) Describe how to estimate the KL divergence $D(p||q)$ using the first-order Von-Mises Expansion approach.

Solution: Let $\widehat{p}$ and $\widehat{q}$ denote kernel density estimators for p and q using the first half of the sample (say we are given n samples from each distribution $\{X_i\}_1^n, \{Y_j\}_{j=1}^n$). The first order Von Mises expansion is:

$$\begin{aligned}
&D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \\
&= \int \widehat{p}(x) \log \frac{\widehat{p}(x)}{\widehat{q}(x)} dx + \int \left(\log \frac{\widehat{p}(x)}{\widehat{q}(x)} + 1 \right) (p(x) - \widehat{p}(x))\, dx + \int \left(-\frac{\widehat{p}(x)}{\widehat{q}(x)} \right) (q(x) - \widehat{q}(x)) dx \\
&+ O\left(\|p - \widehat{p}\|_2^2 + \|q - \widehat{q}\|_2^2 \right) \\
&= \int p(x) \log \frac{\widehat{p}(x)}{\widehat{q}(x)} dx - \int q(x) \frac{\widehat{p}(x)}{\widehat{q}(x)} dx + 1 + O\left(\|p - \widehat{p}\|_2^2 + \|q - \widehat{q}\|_2^2 \right)
\end{aligned}$$

The estimator is based on replacing the two integrals with expectations over the second half of the sample:

$$\widehat{D}(p||q) = 1 + \frac{1}{n/2} \sum_{i=n/2+1}^{n} \log \left(\frac{\widehat{p}(X_i)}{\widehat{q}(X_i)} \right) + \frac{1}{n/2} \sum_{j=n/2+1}^{n} \left(\frac{\widehat{p}(Y_j)}{\widehat{q}(Y_j)} \right)$$

3. **Submodular Feature Selection** Here we study the problem of trying to predict a random variable Z given a collection of random variables $X_1, \ldots, X_p$ (called features). The goal of **feature selection** is to find a small subset of the features that predict Z well.

(a) Show that the mutual information functional $f(S) = I(Z; X_s, s \in S)$ is *not* submodular. This provides evidence that greedy maximization of the mutual information functional may not be a good way to do feature selection.

Solution: Many solutions are possible and we give just one example. Consider the following set of four random variables X_1, X_2, X_3 are bernoulli with probability $p = 1/2$ and $Z = \mathbf{1}[X_1 = X_2]$. Notice that Z is independent of X_3. Notice also that marginally Z is bernoulli with probability $1/2$, but Z is also uniform bernoulli conditioned on either of X_1 or X_2. In particular $p(Z = a, X_j = b, X_3 = c) = 1/8$ for $j = 1, 2$ and for $a, b, c \in \{0, 1\}$. The following are immediate:

$$\begin{aligned}
I(Z; X_3) &= 0 \\
I(Z; (X_1, X_3)) = I(Z; (X_2, X_3)) &= 0 \\
I(Z; (X_1, X_2, X_3)) = H(Z) - H(Z|X_1, X_2, X_3) &= \log 2
\end{aligned}$$

Therefore:

$$I(Z;(X_1,X_3)) - I(Z;X_3) = 0 < I(Z;(X_1,X_2,X_3)) - I(Z;(X_1,X_3)) = \log 2 = 1\text{bit}.$$

which shows that the functional is not submodular.

(b) Show that in the naive bayes model, greedy maximization of mutual information is possible. The naive bayes model posits that $X_i \perp X_j | Z$ for all $i \neq j$ so the distribution factors as $P(Z, X_1, \ldots, X_p) = P(Z) \prod_{i=1}^{p} P(X_i|Z)$.

Solution: We need to show that the mutual information functional is submodular in this case. Using the independence properties of the naive bayes model we have:

$$I(Z;X_S) = H(X_S) - H(X_S|Z) - H(X_S) - \sum_{i \in S} H(X_i|Z)$$

Let $S \subset [p]$ be any subset of the features, and let $i, j \notin S$.

$$\begin{aligned} I(Z;X_S,X_i) - I(Z;X_S) &= H(X_S,X_i) - H(X_S) - H(X_i|Z) \\ I(Z;X_S,X_i,X_j) - I(Z;X_S,X_j) &= H(X_S,X_i,X_j) - H(X_S,X_j) - H(X_i|Z) \end{aligned}$$

This last equality uses the Naive-Bayes assumption, that X_i and X_j are independent conditioned on Z. The difference between the two of these is:

$$\begin{aligned} &I(Z;X_S,X_i) - I(Z;X_S) - (I(Z;X_S,X_i,X_j) - I(Z;X_S,X_j)) \\ &= H(X_S,X_i) - H(X_S) - H(X_S,X_i,X_j) + H(X_S,X_j) \\ &= H(X_i|X_S) - H(X_i|X_j,X_S) \geq 0 \end{aligned}$$

The last inequality follows since conditioning does not reduce entropy. Since this holds for any S, i, j, this shows that the functional is submodular.